DOCUMENT RESUME

ED 411 301 TM 027 329

AUTHOR Gitomer, Drew H.
TITLE Challenges for Scoring Performance Assessments in the NBPTS System.
PUB DATE 1997-03-00
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Higher Education; Licensing Examinations (Professions); *Performance Based Assessment; *Scoring; *Standards; *Teacher Certification; Teacher Evaluation; *Test Construction; Test Content; Test Use
IDENTIFIERS *National Board for Professional Teaching Standards

ABSTRACT

Assessing complex teaching performance in the National Board of Professional Teaching Standards (NBPTS) has caused the Educational Testing Service to wrestle with fundamental scoring issues that are both conceptual and technical. This report reviews the challenges encountered, how they are being addressed, and what the NBPTS effort has learned over the past 3 years. Scoring a performance assessment is the overarching consideration in developing an assessment. In current development work for the NBPTS, scoring is the framework for the design of tasks and the entire assessment. The following challenges have been faced in developing assessment and scoring: (1) defining scoring against task-independent standards; (2) handling a broad range of content and contexts; (3) resolving pedogogical, cultural, and contextual bias; (4) interpreting unfamiliar representational forms and content; and (5) defining "The Standard" in setting the cut score. The NBPTS scoring system has evolved dramatically over the last 3 years into a process that makes certification decisions much fairer and more defensible. Without a coherent design process that begins with a clear sense of the desired claims to be made about a candidate, scoring will be found lacking. (Contains four figures.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 411 301

Challenges for Scoring Performance Assessments in the NBPTS System

Drew H. Gitomer
Educational Testing Service

Paper presented at Annual Meeting of the American Educational Research Association, March, 1997, Chicago

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Drew Gitomer
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM027329

BEST COPY AVAILABLE

Challenges for Scoring Performance Assessments in the NBPTS System

Drew H. Gitomer
Educational Testing Service

Assessing complex teaching performance in the NBPTS effort has compelled us to grapple with fundamental scoring issues that are both conceptual and technical. In this talk, I would like to overview challenges we have encountered, how we are addressing those challenges, and what we have learned over the past three years.

The structure of the NBPTS assessments is discussed in the paper by Mari Pearlman in this section. We have come to understand that the scoring of a complex performance assessment is the overarching consideration in assessment development. This is in stark contrast to earlier NBPTS attempts and much of the work in performance assessment in general, in which interesting and provocative tasks are the primary consideration. In our current development work for the National Board, scoring is the framework for the design of tasks and the entire assessment, because the scoring process is the source of claims about whether or not candidates demonstrate the qualities that define highly accomplished teaching. Therefore, scoring considerations shape task design - it is not something that is defined subsequent to task design (see Gitomer & Steinberg, in press).

Our scoring system had to accommodate a set of challenges shown in Figure 1. It is worth noting that these challenges are not unique to NBPTS

assessment. Any complex performance, in which the assessment tasks mandate quite unique responses, would face these same set of challenges. These types of assessments stand in contrast to much more constrained tasks for which individuals are to respond in a highly specified manner, and for which scoring criteria are narrowly defined.

In essence, every assessment system faces a tradeoff in how it disciplines the scoring process. At one extreme, a multiple-choice test imposes very rigid discipline on the responses of those being assessed. Virtually no interpretation of responses needs to be made other than scoring each response as correct or not. At the other extreme, assessments such as those of the NBPTS, allow for significant variability in the responses. For these assessments, the discipline of interpretation must be imposed in the scoring structures. Such discipline results from processes of social moderation and judgment that occur throughout the scoring process.

Challenges for NBPTS Scoring
1. Defining scoring against task-independent standards
2. Handling a broad range of content and contexts
3. Resolving pedagogical, cultural, and contextual bias
4. Interpreting unfamiliar representational forms and content
5. Defining "The Standard" - Setting the cut score

Figure 1.

Defining scoring against task-independent standards

The NBPTS assessment tasks are interpretations of standards. The standards are not prescriptive of assessment tasks, nor should they be. Further, each task is assessing the coordination of teaching that cuts across multiple standards. The challenge for scoring is to create a mapping between the assessment and the standards that honors the intent of the standards and also has sufficient specificity so that assessors can recognize how the aspects of the standards are embodied in the specific assessment task.

Standards documents are inherently flat. In the written enumeration of a set of standards heuristic distinctions are made among attributes that are necessarily intertwined. Further, everything contained within the standards is deemed important - seldom is priority given to one standard, or aspect of a standard, at the expense of any other. Scoring however, does require making clear value judgments, deciding which aspects of performance are absolutely critical and which might be desirable, but not absolutely determinant. Scoring also requires an explicit articulation of the ways in which standards are manifest in a complex, integrated performance.

The bridges between the standards and specific assessment tasks are embodied in scoring rubrics for each entry. An example of a rubric for teachers of English/Language Arts is presented in Figure 2. In the rubric, certain aspects of performance are given primacy and are more determinant of a score than are

Figure 2

Analysis of Student Response to Literature Rubric

4

The **4** level performance offers **consistent and convincing** evidence that the teacher has a thorough knowledge of students as individual learners, sets high, worthwhile and attainable goals for student learning, and provides a context for reading that encourages students' active exploration of literature. The **4** level performance offers **consistent and convincing** evidence that the teacher recognizes multiple interpretations and requires them to be grounded in the text, recognizes students' progress, encourages active interpretation and critical reading of literary and non-literary texts, and offers means for students to build on their accomplishments. The **4** level performance offers **consistent** evidence that the teacher employs varied, appropriate instructional resources to support students' growth as readers. There is **consistent** evidence of ongoing assessment of reading growth and of effective communication with students about their responses to literature. The **4** level performance gives evidence that the teacher is able to describe his/her practice accurately and to reflect insightfully on its effectiveness in meeting the challenges of teaching literature.

3

The **3** level performance offers **clear** evidence that the teacher has a thorough knowledge of students as individual learners, sets high, worthwhile and attainable goals for student learning, and provides a context for reading that encourages students' active exploration of literature. The **3** level performance offers **clear** evidence that the teacher recognizes multiple interpretations and requires them to be grounded in the text, recognizes students' progress, encourages active interpretation and critical reading of literary and non-literary texts, and offers means for students to build on their accomplishments. The **3** level performance offers evidence that the teacher employs varied, appropriate instructional resources to support students' growth as readers. However, these resources may be less varied than in a **4** level performance. There is **clear** evidence of ongoing assessment of growth as a reader and of effective communication with students about their responses to literature. However, the assessment and/or communication may be less insightful than in a **4** level performance. The **3** level performance gives evidence that the teacher is able to describe his/her practice accurately and to reflect on its effectiveness in meeting the challenges of teaching literature. **A 3** level performance may show imbalance in the analysis and/or evidence presented for each sample. **One of the samples may be more indicative of accomplished practice than the others, but viewed as a whole there is clear evidence of a 3 level performance.**

2

The **2** level performance offers **limited** evidence that the teacher has a knowledge of students as individual learners. It also exhibits limited evidence that the teacher provides a context for reading that encourages students' exploration of literature. The goals for student learning may be

* vague, trivial, or inappropriate
* clearly unrelated to the instruction.

The 2 level performance offers **limited** evidence that the teacher recognizes multiple interpretations and/or insists they are grounded in text and that the teacher recognizes students' progress. The evidence that the teacher encourages active interpretation and critical reading of literary and non-literary texts and/or evidence that the teacher provides means for students to build on their accomplishments is limited or missing. Instructional resources and activities may be inappropriate, formulaic, and/or lacking a plausible rationale. There is **limited** evidence of assessment of growth as a reader and/or evidence of ineffective communication with students about their responses to literature. The 2 level performance gives **limited** evidence that the teacher is able to describe his/her practice accurately. The reflection is weak or skeletal and includes **limited or no** evidence of meeting the challenges of teaching literature. **In general, the 2 level performance is characterized by evidence that may hint at accomplished practice, but is too fragmented or uneven to support a clear classification as a 3 level performance.**

1

The 1 level performance offers **very limited or no** evidence that the teacher has a knowledge of students as individual learners and provides a context for reading that encourages students' participation. The goals for student learning may not be goals at all, but rather activities. Goals, when stated, are trivial, vague, or inappropriate. The 1 level performance offers **very limited or no** evidence that the teacher recognizes multiple interpretations and/or insists they are grounded in the text and/or recognizes students' progress. The evidence that the teacher encourages active interpretation and critical reading of literary and non-literary texts and /or evidence that the teacher provides means for students to build on their accomplishments is **very limited or missing**. Instructional resources and activities may be inappropriate, unrelated to goals, and/or lacking rationale. There is **very limited or no** evidence of assessment of growth as readers and/or effective communication with students about their responses to literature. The 1 level performance gives **very limited or no** evidence that the teacher is able to describe his/her practice accurately. The reflection is missing or unconnected to the instructional evidence.

BEST COPY AVAILABLE

other aspects of performance. The rubric also takes the generally broad language of the standards and translates it to the specific requirements of an entry. During assessor training, the bulk of the time is devoted to anchoring the meaning of such terms as "convincing" and "consistent," and "limited" and "plausible" by leading assessors through multiple examples of teachers' responses that embody the essential elements of such characteristics, albeit in different guises (e.g., classroom contexts, styles, settings, etc.)

Handling a broad range of content and contexts

The portfolio entries, and to a lesser extent the assessment center exercises, encourage a broad range of responses. In portfolio entries, the challenge for candidates is to provide evidence of meeting the standards through their own practice. The entries do not prescribe particular teaching methods, or even content. Asking candidates to show how they meet the standards by, for example, encouraging productive discourse in the classroom about an important idea, or demonstrating their assessment practices, leads to teachers offering evidence of their practice that differs on a significant number of levels.

For one, each certificate encompasses a wide range of teaching circumstances. For example, candidates for high school mathematics might be teaching AP Calculus, but they also might be teaching a remedial algebra course.

Scoring a teacher's understanding of content and its pedagogy cannot depend on the mathematical sophistication of the course content or its students.

Second, teaching occurs in many different contexts, even for an individual teacher. Aside from the level of the course, one class may contain fifteen students, another thirty or more. Some schools have significantly more material resources at their disposal than others. Some classes may not have a socially or academically diverse makeup, while others are extremely diverse. Finally, teachers may be teaching in relatively homogenous classrooms, but some will be teaching well-to-do white children and others will be teaching African-American children from economically impoverished homes.

For all these differences, it is the standards that create the common conceptual structure against which so many different kinds of performance can be considered. Teachers' portfolios can not be judged on the basis of context or course content - those are the cards that are dealt. As assessors, we can legitimately ask, though, whether portfolio entries provide evidence of meeting specified standards, *given* the teaching context, not *independent* of the teaching context.

Note the distinction between acknowledging teaching context and acting as if differences didn't exist. They do, and assessors must be able to recognize and attune their scoring in response. Judging content knowledge for the calculus and remedial algebra course cannot be made on the basis of our more familiar

orderings of content depth. We must ask whether, given a specific context, does the teacher demonstrate practice that is aligned with the standards.

Resolving pedagogical, cultural, and contextual bias

Not only must scoring acknowledge differences in teaching that are orthogonal to the standards, but assessors must avoid the bias of giving differential consideration to features of a portfolio entry that may be familiar in terms of the assessor's own teaching. Assessors tend to be highly accomplished teachers in their own right. We want assessors to focus on the issue of whether the portfolio entry shows evidence of meeting the standards, not whether the entry represents teaching to which they are personally sympathetic.

The teaching of language arts is a domain in which there are very strong and very different views on how to teach writing, for example. Some teachers are strong proponents of a writer's workshop approach, while others pursue more traditional approaches. Scores cannot be given on the pedagogical approach that the teacher adopts, but on the basis of how the teacher uses a given approach to help students develop important skills in and understandings about writing, and how conscious, deliberate, and thoughtful the teacher's rationale for the approach is, given the context in which the teaching takes place. This is especially important if the standards and the assessments are to maintain

a currency that extends beyond particular pedagogical approaches that become more or less popular over time.

In order for assessors to be able to interpret teacher entries that reflect a range of content, contexts and pedagogical strategies, assessor training involves a great deal of exposure to different examples of performance. Assessors spend up to 4 days learning to search for evidence and apply a rubric to a given entry. The direct challenge is for assessor training to help assessors see that performance level is not confounded with content, context, and pedagogical approach. During training, assessors are presented with exemplar entries that illustrate different contexts, alternative pedagogical approaches, and different quality of performance. Training is designed to explicitly challenge any stereotypes or preconceptions that assessors may bring to this task.

Interpreting unfamiliar representational forms and content

In the portfolio, teaching practice is assessed through the examination of classroom artifacts and a teacher's written commentary. Artifacts include videotaped segments of classroom discourse, samples of student work, classroom assessments, and instructional materials. These forms of evidence are, for most teachers, alien to any evaluation process. More typically, teachers are evaluated on the basis of a very occasional classroom observation.

In order for National Board assessors to make complex inferences about teaching accomplishment by examining artifacts and teacher commentary, they must be able to grapple with forms of evidence that have never been associated with formal characterizations of teaching quality. Therefore, a scoring system must help assessors make connections between classroom artifacts and teacher commentary to claims about teacher accomplishment. Fundamentally, the National Board assessment system subscribes to the belief that products of classrooms, such as videotapes and student work, are powerful and valid forms of evidence for making claims about teaching practice. Because that belief is not accepted wisdom, assessor training requires a great deal of attention to establishing these connections. In order to make such connections, assessors devote training time to learning how to consider classroom and teacher-produced evidence in terms of quality, consistency, and clarity.

Assessor training is designed to teach assessors how to search for evidence in candidate entries, and to make inferences about the evidence with respect to standards. In Figure 3 are guiding questions that assessors are asked to consider for one entry. These questions are designed to help assessors construct the connections between the evidence submitted by the candidate and the rubric. The questions illustrate how assessors learn to consider each piece of the entry separately as well as together. Always, assessors are evaluating the coherence among various pieces of submitted evidence. Lack of coherence

Figure 3

Guiding Questions - Analysis of Student Response to Literature

For Students A, B, and C: (Analyze each commentary and folder separately. Record evidence as you read. Cite examples.)

Analysis Form

1. What is the nature of the evidence the candidate provides about
   a. students
   b. instruction
   c. connections made between the information about students and the practice
   d. goals of instruction
   e. connection between the goals and the assignment(s)/prompt(s)

2. What is the evidence that the candidate accurately and insightfully assesses the following:
   a. student response in light of instructional goals
   b. student response in relation to literature instruction that fosters individual growth as a reader

3. What is the evidence that the candidate understands the role of feedback in building readers' abilities and provides vehicles for effective feedback to these readers?

Analysis and Student Work Together

4. What is the nature of the "fit" between what the teacher says and what the teacher does?

5. What are the ways in which the student's work explains the analysis and/or the analysis explains the student's work? Be specific. Remember, the two sources of evidence can support and enhance each other or conflict and undermine each other.
   - design and execution of instructional goals
   - influence of the student's work on future instruction
   - specific aspects of the student's response that demonstrate growth as a reader
   - assessment of feedback to the student

Reflective Essay

6. What is the nature of the evidence that the candidate provides that explains his/her challenges and goals for teaching literature in light of the students' work presented?

Before assigning the final score:

Overall, considering
- the evidence of all three commentaries
- the responses of all three students
- the reflective essay

What is the evidence of this candidate's command of literature instruction as such instruction is delineated in the EA/ELA standards?

BEST COPY AVAILABLE

means that the story doesn't quite gel, and makes for a less compelling argument about the candidate's level of accomplishment.

Assessors are also provided with structures for looking at evidence. Figure 4 presents the scoring path for examining the evidence in an entry. Scoring paths not only are helpful to assessors, but they serve to standardize the examination process so that all candidates are assured that their work is looked at under similar conditions and constraints.

Finally, assessor training is dedicated to helping assessors create a trail of evidence for each candidate's response. Assessors record specific evidence, making explicit connections to the rubric. An example of such an evidence sheet, completed, is presented in Figure 5. The evidence sheet is organized according to specific parts of the entry, but the contents of the evidence sheet are clearly grounded in aspects of the rubric. Assessors use the Guiding Questions as the structure for their records of evidence, responding to each as they move through the entry.

Defining "The Standard" - Setting the cut score

The paper by Charlie Lewis and Mari Pearlman (also in this session) describes in detail the standard setting process and some of the ways we have addressed this issue. Essentially, the challenge for the assessment system is to make a single decision, to certify or not. The current system scores each of ten entries separately, with a unique set of assessors and trainers. The challenge is to

Figure 4

Scoring Path for Analysis of Student Response to Literature

At the start of each scoring session

1. Review the standards addressed in the exercise.
   Standard I Knowledge of Students
   Standard II Curricular Choices
   Standard VI Reading
   Standard XI Assessment
   Standard XII Self-Reflection

2. Review the directions for the exercise.

As you score each candidate

1. Review the rubric for scoring the exercise.

2. Review the guiding questions and use them to document the evidence for the exercise.

3. Read the analysis and submitted materials for Student A and document the evidence on the evidence sheet.

4. Read the analysis and submitted materials for Student B and document the evidence on the evidence sheet.

5. Read the analysis and submitted materials for Student C and document the evidence on the evidence sheet.

6. Read the reflective essay and record the evidence on your evidence sheet.

7. Complete the overall summary using "Before assigning the final score" as your guide.

8. Using the rubric and the evidence, assign a score.

Figure 5

Candidate ID: 12 Score: 2 Assessor ID: _____

Analysis of Student Response to Literature Evidence Summary Sheet

Student A	Student B	Student C
1/2 Knowledge of students - General class info. - Limited - Lack of prior knowledge Instruction - Goals - Worthwhile - Not esp high (evidence from play) - FCA - Vocab, C/E, detail C/E doesn't fit assignment · Sequence / Activities - Reflect process · Resources - Appropriate · Student growth as reader - Thinking, questioning underwent change - Vague, limited evidence 3. Feedback - Notes, conference · Assessment - Add text evidence - Build on reader · Transition - How? Future inst. - Limited as to how to build - Formulaic 4/5. "Fit" ok - Response helps build growth · Future Inst. - Implied area - Details FCA - C/E - Doesn't fit - Goal - No alignment to activity	1/2 Knowledge of Student - Limited evidence - "Sue" Instruction - Goals - Worthwhile, attainable - Compare / Contrast / Organize Resources - Appropriate · Sequences - OK - Not convincing · Activities - Appropriate · Growth as reader - No real evidence - Needs detail (Evidence as writer) 3 Feedback - Through process - Conference · Next Steps - General plan - "Put more time" · Limited Ev. - Read other essays to "see" Characters 4/5 - "Fit" - OK · Response - Vague - "Conclusion usual" Limited response · Next step aligns to work	1/2 Knowledge of Student - Clear evidence Instruction - Goals Worthwhile, attainable - Character Analysis · Resources - Appropriate · Sequence / Activities - Appropriate - Limited Growth as reader - Limited evidence - "Strong development and inquiring mind..." · Weak analysis of salient features 3. Feedback - Discussion w/ student · Prompts student to reflect - · Future Inst. - Limited evidence - Read extensively, produce better essays 4/5 "Fit" - OK · Response is weak, limited · Next step aligns to work
Reflection Goals and themes are broad - Rationale for theme is clear, could be more convincing. Limited evidence of challenge and these 3 students in context of teaching. "Movement and growth..." - Does not reflect on practice teaching to this		
Overall Evidence of exploration and appropriate resources. Teacher recognizes progress, but shows limited evidence of response / instruction to help build on it. Formulaic approach - Read more... all do same assignment. Weak reflection on practice and how it relates to students and progress. Hints at accomplished practice. It's uneven. Mixes goals and activities		

a/assfev.doc 7/16/96

AVAILABLE

combine these ten discrete pieces of information to make the certification decision.

The combining of information has forced us to challenge some of our deeply held measurement beliefs and practices. First, both the standards and our understanding of teaching make it clear that there is no theoretical justification for assuming that all entries are random samples of a construct called "accomplished teaching." Instead, we are taking unique slices of a complex performance that, by design, have differing amounts of theoretical overlap with each other.

In deciding how to combine these unique slices, a number of policy decisions have been needed. For one, a major consideration has been to consider scores as compensatory. Though some certification systems require candidates to meet a certain standard for each component of the assessment, National Board assessment allows candidates to compensate for some low scores with other high scores.

A second major issue concerns how to weight each of the entries in the aggregation of evidence for or against certification. A variety of methods have been used, all leading to essentially the same pattern of weightings - portfolio entries are considered as very important, while assessment center entries are considered less so. This is not surprising, given that portfolio entries require the coordination of so many aspects of teaching integral to the standards.

A critical feature of the assessment system is that we are able to query the standard setting process and results quite intensively. First, individual entry scores are established independent of any overall consideration of teaching accomplishment. The combining of scores is done mathematically and gauged against the standard. However, as Lewis and Pearlman, and Bond demonstrate, we can, and do, return to the original data to determine whether the overall certification decision is defensible when considering the total original data produced by the candidate. The willingness to return to the original data allows us to question assumptions of any aggregation and weighting models employed to make the certification decision.

Conclusions

The NBPTS scoring system has evolved dramatically over the last three years. We believe that the scoring process that has developed makes the certification decision much fairer and more defensible. Other papers in this session will address a variety of indicators of such measurement quality. Ultimately though, the quality of the scoring process is bounded by the quality of the underlying standards and the quality of the assessment's design. Without a coherent design process that begins with a clear sense of the desired claims to be made about a candidate, scoring will be found lacking, no matter how elegant the scoring infrastructure. We believe that our attention to the link between

scoring and design has resulted in an assessment system that can support the high-stakes certification decisions that are the linchpin of the National Board system.